TIPOLOGIA I CICLE DE VIDA DE LES DADES

PRÀCTICA 1

Integrants del grup: Pere Moles i Cristina Vilageliu

1. Context

El FC Barcelona és una de les entitats amb més rellevància en el nostre país i és per això que els resultats dels partits de les diferents disciplines que s'engloben dins del Barça, els nous i veterans jugadors i els resultats dels diferents tornejos on participen son comunament utilitzats per a fer datasets i utilitzar aquestes dades per a fer informes analítics de rendiment, històrics, etc. En el nostre cas, ens hem volgut situar en la posició de l'aficionat que va al camp o a l'estadi per a tractar un altre aspecte important: quant val anar al camp? Per el nostre projecte hem decidit fer *web scrapping* de la pàgina web oficial del FC Barcelona.

L'adreça web utilitzada ha sigut la següent: https://www.fcbarcelona.es/

La idea és aconseguir un dataset mitjançant *web scraping* amb informació sobre els preus de les entrades dels propers partits de les diferents disciplines que s'engloben dins del Barça. D'aquesta manera, pretenem trobar respostes a quines son les entrades més econòmiques, quines les més cares, quina és la secció de l'estadi amb millor preu, o quina disciplina s'ajusta més al nostre pressupost, etc.

2. Títol

El títol que proposem per al data set és: Preus de les Entrades del FC Barcelona per disciplina, secció i rival. El què ens interessa del nostre estudi és el preu de les entrades per a poder comprar-lo. Aquest preu varia segons la disposició del seient a la graderia i del rival contra el qual juguin. Per això, ens interessa que en el títol del nostra dataset quedin reflectides les parts importants que influeixen en la dada

d'interès: quin preu tenen les entrades d'una disciplina particular, en la secció de l'estadi que ens interessa contra el rival que volem veure?.

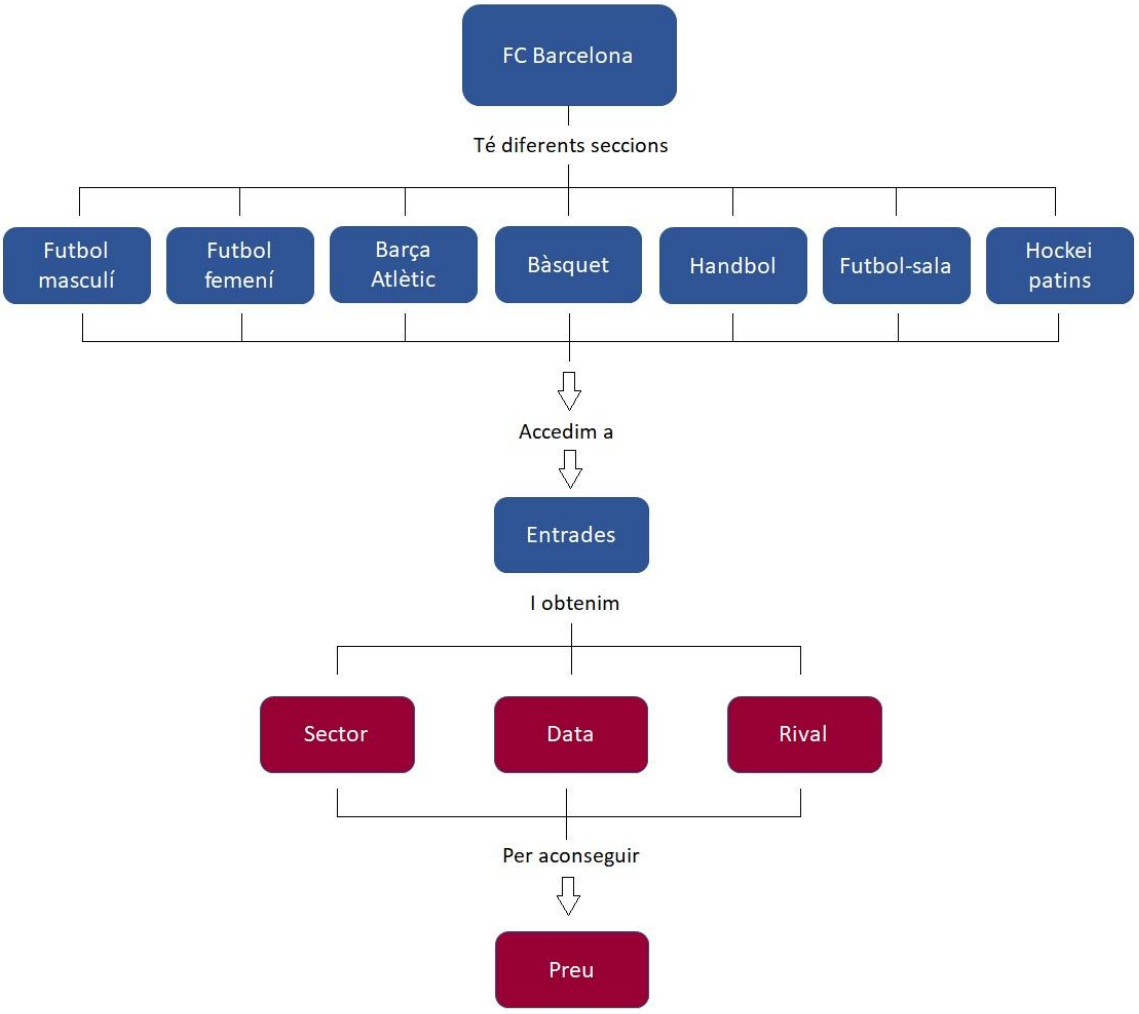
3. Descripció del dataset

El FC Barcelona consta de diferents disciplines o seccions (futbol masculí i femení, bàsquet, handbol, etc.) i per a cada una d'elles en volem aconseguir el preu de les entrades. Per tant, el nostre dataset consta d'un índex i 4 variables: el sector de l'estadi on es troba el seient, la data del partit, el rival contra el qual juguen i el preu de l'entrada. Les seccions de l'estadi apareixen una sola vegada en el dataset amb el preu general de l'entrada per als seients d'aquella secció.

Serà interessant veure i analitzar quina de les variables influeix més sobre el preu i també veure, si el projecte s'allargués en el temps, l'evolució del preu de les entrades durant varis anys contra els mateixos rivals per a veure quina tendència tenen.

4. Representació gràfica

A continuació veurem en un gràfic il·lustratiu un resum del nostre projecte per entendre-ho millor:



Partim de l'entitat del FC Barcelona ja que és l'objecte del nostre estudi i també la pàgina web sobre la que hem fet *web scraping* per a aconseguir el dataset.

El FC Barcelona, com ja hem comentat, té diferents seccions o disciplines en què els aficionats poden comprar entrades per a veure els partits. En total son 7 les disciplines que engloba el Barça: Futbol masculí, Futbol femení, Barça Atlètic, Bàsquet, Handbol, Futbol-Sala i Hockei patins.

Per a cada disciplina ens interessa accedir a les entrades. En el nostre cas, a través de rastrejadors o *scrapers* hi accedim literalment seguint l'enllaç de la pàgina web. De cada entrada n'obtenim el sector de l'estadi o graderia on es troba el seient, la data del partit i el rival contra el qual juga i aconseguim el preu per a cada cas.

5. Contingut

El nostre dataset està format per la informació disponible a la pàgina web del FC Barcelona sobre les entrades als partits de les diferents disciplines que formen part del Barça. En la pàgina web només estan disponibles les entrades per als partits de la temporada actual, és a dir, el nostre dataset només recull informació de la temporada 2022-2023. Aquesta situació és normal, ja que encara no és possible saber, ni per el propi club, com serà la planificació de partits de la temporada següent.

Com hem anat comentant, el nostre dataset consta de quatre variables: Secció, Rival, Dia i Preu.

La variable Secció és una variable categòrica i ens indica en quina part de la graderia es troba el seient. En el cas de la disciplina de futbol trobem per a aquesta variable objectes com “Corner 1a”, “Corner 2a”, “Gol 1a”, “Gol 2a Superior”, etc. Per a cada secció de l’estadi tindrem un preu, així que per a cada disciplina la secció de la graderia apareix una única vegada.

La variable “Rival” és una variable categòrica que ens indica contra quin rival juga el Barça. Aquesta dada serà important per a fer anàlisis ja que depenent del rival pot ser que els preus canviïn i podríem fer un estudi sobre quin rival és més car anar a veure quan ens visita. Exemples de rivals poden ser “Espanyol”, “Real Sociedad”, “Cazoo Baskonia”, etc.

La secció “Dia” és una variable temporal que ens indica quin dia està planificat jugar el partit.

Per últim, la variable “Preu” és una variable numèrica que ens indica el preu de les entrades per a cada secció de la graderia, contra un rival específic en el dia que s’enfronten.

Un dels problemes amb els quals ens trobem en el nostre dataset és la necessitat d’actualitzar les dades. Si seguim només les entrades a partits de lliga de les diferents disciplines no tindríem cap problema ja que aquests partits es planifiquen a principis de temporada i, malgrat no hi ha hagi cap incidència, no es modifiquen les

dates ni els rivals ni, esperem, els preus. Però en altres competicions eliminatòries, com pot ser la Champions de futbol o la Copa del Rei de Bàsquet, el calendari de partits s'actualitzaria setmanalment depenent de si el Barça passa de fase o no i això ens obligaria a mantenir la base de dades actualitzada setmanalment.

6. Propietari

Fent recerca en la pàgina web del FC Barcelona, trobem que totes les dades que es troben a la pàgina, incloent-hi textos, imatges, logos, codis font i més, pertanyen al CLUB (és a dir, al Futbol Club Barcelona) així que ells són els propietaris de les dades. Aquí deixem l'extracte de la pàgina on ens informa:

"Els textos, imatges, logos, signes distintius, sons, animacions, vídeos, codis font i resta de continguts inclosos a la web són propietat del CLUB, que en disposa al seu cas del dret d'ús i explotació i, en tal sentit, es converteixen en obres protegides per la legislació aplicable en matèria de propietat intel•lectual i industrial, nacional i internacional vigent en cada moment. (...)".

Els punts en els què ens hem centrat per a assegurar-nos de que complim amb els aspectes legals i ètics de la pàgina web son:

- Rastrejar només informació pública: Les dades que apareixen en el nostre dataset són accessibles per al públic a través de la pàgina oficial del FC Barcelona i no cal registrar-se ni ser soci per a poder veure-la. En el nostre dataset lo únic que hem fet és ajuntar tota aquesta informació de forma compacta de manera que en un sol arxiu podem trobar informació de totes les entrades de tots els partits disponibles de totes les disciplines. L'ús de *web scraping* en el nostre projecte és moderat.

- No causar dany: Com hem dit en el punt anterior, tota la informació que hem recopilat és accessible des de la pàgina web oficial de l'entitat. No hem sobrecarregat el servidor amb moltes peticions ja que el nostre punt d'interès és

el preu de les entrades i no hem tingut la necessitat d'accedir a servidors als quals no tenim accés.

- Utilitzar la informació de manera justa: En el nostre cas, la finalitat d'aquest projecte és pròpiament didàctica i no tenim cap intenció d'utilitzar les dades amb una finalitat comercial o treure'n cap benefici econòmic. A més, recordem que aquestes dades pertanyen al CLUB (FC Barcelona).

Fent recerca, no hem trobat cap cas d'estudi amb dades com les nostres, és a dir, que recopilin la informació dels preus de les entrades durant tota una temporada, però si n'hem trobat de similars.

El cas d'estudi plantejat per Pérez González, Fernandez-Luna, de la Riva i Burillo titulat "Poder adquisitivo y precio de entradas de fútbol en España ¿Es más o menos caro el fútbol ahora que hace 50 años? El caso del club Atlético de Madrid" és un exemple d'un projecte similar.

En aquest projecte, fan una comparació del preu de les entrades de l'estadi Vicente Calderón, seu futbolística de l'Athletic de Madrid. La comparació és el preu de l'entrada l'any que es va inaugurar l'estadi, el 1966, amb el preu de l'entrada l'any 2016. Per a fer-ho, han utilitzat deflactors i el PIB per càpita dels dos anys per a estimar la possibilitat que una persona de classe mitja pugui visitar el camp de l'Athletic de Madrid.

La conclusió a la qual han arribat en l'estudi és que tot i que el preu de les entrades s'ha triplicat en 50 anys, també s'ha triplicat el PIB per càpita i el preu d'entrades en tot el sector cultural. És a dir, per a un aficionat de classe mitja la possibilitat econòmica d'anar al camp a veure el seu equip és la mateixa al 2016 com ho era en el 1966.

L'enllaç al cas d'estudi que acabem de comentar és el següent:

(PDF) Poder adquisitivo y precio de entradas de fútbol en España ¿Es más o menos caro el fútbol ahora que hace 50 años? El caso del club Atlético de Madrid (researchgate.net)

Navegant per internet trobem moltes pàgines web que contenen bases de dades molt interessants sobre el futbol, per exemple, així com també d'altres esports. Aquestes pàgines web es centren en l'aspecte esportiu dels clubs i fan estadístiques amb històrics i inclús prediccions de partits. Una d'aquestes pàgines que recull dades històriques de clubs de futbol és BDFutbol (Datos Históricos y Estadísticas de la Liga Española de Fútbol, Premier League, Serie A, Bundesliga, Ligue 1, Primeira Liga y Eredivisie (bdfutbol.com)). A BDFutbol es poden trobar dades de totes les competicions, clubs, jugadors, entrenadors, àrbitres, seleccions i son dades històriques, actuals, estadístiques i més. Altres exemples d'aquests tipus de pàgines que emmagatzemen dades esportives son TransferMarkt (Football transfers, rumours, market values and news | Transfermarkt) o WhoScored (Football Statistics | Football Live Scores | WhoScored.com).

És difícil comparar el nostre projecte amb aquestes pàgines web ja que l'enfocament és molt diferent i la manera de recollir la informació també. De totes maneres, com veurem en el següent apartat, podem aprofitar la importància que també tenen aquestes dades i plantejar-nos d'afegir-les al nostre dataset per a poder respondre més preguntes sobre la nostra variable d'interès, el preu de les entrades.

7. Inspiració

La nostra inspiració per a realitzar aquest projecte ha sigut la nostra pròpia experiència a l'hora de comprar entrades. El què pretenem amb aquest dataset és poder aconseguir les millors entrades, en el sector de l'estadi que més ens agradi i al millor preu. Gràcies a la recol·lecció de informació que hem dut a terme, també podem respondre a més preguntes com per exemple:

- Quina és l'entrada més cara de la temporada? A quina disciplina i secció de l'estadi pertany? Contra quin rival és?
- Contra quin rival són més cares les entrades segons la disciplina?
- Contra quin rival són més cares les entrades segons la disciplina?
- Quina és la disciplina més assequible per anar a veure un partit?
- Quina és la disciplina més cara per anar a veure un partit?
- Quina secció de l'estadi és més cara?
- Quina secció de l'estadi és més econòmica?
- De cada rival quina és la secció més econòmica?
- Influeix el dia en què es juga en el preu de les entrades?

Anem a relacionar les preguntes que podem respondre gràcies al nostre dataset amb les casos d'estudi esmentats en l'apartat anterior.

En el cas del projecte en el què comparen els preus de les entrades entre dos partits amb 50 anys de diferència, ens podria inspirar a preguntar-nos les diferències en els preus entre un partit i un altre depenent del rival. Si obtinguéssim aquestes dades cada any, podríem fer un anàlisi històric i observar si el preu de les entrades pugen o baixen comparant els partits amb el mateix rival, per exemple. També podríem observar, en línia general, si amb el temps les entrades s'encareixen o si esdevenen més econòmiques.

Respecte a les pàgines web amb informació estadística dels partits, jugadors, etc. podem plantejar-nos ajuntar aquesta informació amb les dades del nostre dataset i així poder fer-nos preguntes relacionant els jugadors que han participat en el partit, si l'any anterior van guanyar o perdre contra el rival i més i veure si influeix en el preu de l'entrada.

8. Llicència

La llicència amb la que publicaríem la base de dades resultant és Released Under CC BY-NC-SA 4.0 License.

Aquesta llicència permet compartir el dataset, copiar i distribuir les dades en qualsevol mitjà o format, i adaptar-lo, transformar-lo i crear noves variables a partir de les que ja tenim.

El motiu principal per el qual hem seleccionat aquesta llicència és perquè el seu propi nom ens indica que no pot ser usada per a fins comercials (NC – NonComercial). Aquest és un dels termes que imposa la llicència així com sempre donar crèdit a l'autor, a la llicència i indicar si s'ha realitzat algun canvi. L'últim terme que s'ha de garantir sota aquesta llicència és que si s'adapta, transforma i es creen noves dades, s'ha de mantenir aquesta llicència.

9. Codi

El codi l'hem estructurat de la següent manera:

- Dues funcions auxiliars __text_cleaner i __date_cleaner que corregeixen alguns errors tipogràfics que hem trobat al scrapejar text de la web, i passar les dates al format estàndard de Datetime de Python.
- La funció scrap_matches_url és una funció que donada la pàgina de navegació d'una secció, per exemple futbol, retorna un dataframe amb la llista de partits, horaris, preus i seients de cada partit. Originalment voliem obtenir els enllaços de cada pàgina d'entrada utilitzant simplement BeautifulSoup i obtenir l'href de l'enllaç per fer una request d'aquesta pàgina. Però ens vam trobar que aquest enllaç no era el definitiu, i redirigia a una pàgina prèvia per evitar robots. Per això vam decidir utilitzar Selenium, per simular el click al botó de les entrades.

- Aquest és l'objectiu de la funció __selenium_enter_ticket_page. L'objectiu principal és retornar la url de la pàgina de les entrades d'un partit concret. Aquesta funció té bastantes excepcions depenent de l'esport, ja que ens vam trobar que alguns esports tenien més d'una pàgina abans d'arribar a la definitiva, depenent de l'esport tenia anuncis publicitaris diferents, fins i tot el nom de tags de l'html de la pàgina tot i ser similars tenien noms diferents…Per tot això hi ha diversos casos i excepcions.
- La funció __scrap_prices és la funció que un cop dins de la pàgina d'entrades, extreu els diferents preus i localitats a l'estadi, i retorna un dataframe amb aquesta informació.

10. Dataset

https://zenodo.org/record/7348855#.Y301YNJBxH5

11. Video

https://drive.google.com/file/d/1KP-YhfLh39FarFL3zrhr2d52W8qg3B5g/view?usp=sharing

Taula de contribucions

Contribucions	**Signatura**
Investigació prèvia	CV / PM
Redacció de les respostes	CV / PM

Desenvolupament del codi	PM
Participació al video	CV / PM